DOCUMENT RESUME

ED 336 420 TM 017 217

AUTHOR Ligon, Glynn; Mangino, Evangelina
TITLE A Call for a New National Norming Methodology.
INSTITUTION Austin Independent School District, Tex. Office of Research and Evaluation.
PUB DATE Apr 91
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; Computer Uses in Education; *Data Analysis; Data Collection; Educational Change; Educational Innovation; *Educational Technology; Elementary Secondary Education; Evaluation Methods; *National Norms; Sampling; Scoring; *Standardized Tests; *Test Norms

ABSTRACT

Issues related to achieving adequate national norms are reviewed, and a new methodology is proposed that would work to provide a true measure of national achievement levels on an annual basis and would enable reporting results in current-year norms. Statistical methodology and technology could combine to create a national norming process that would publish an annual national norm soon after test users complete their regular annual testing programs. This approach would be supplemented by a small national normative sample and adjusted by a factor to reduce the influence of users. This approach would also include almost immediate turnaround of the current-year norms to allow schools to report their annual scores using those norms. Schools already giving a test would transmit their data electronically to a central location. Schools in the normative sample would test and transmit their data. A current year norm table could then be produced. This type of norming procedure would probably require a national center for test norming, as well as cooperation among test publishers. Cooperation from school districts would be secured by the fact that they could have national norms within weeks of transmitting their raw data. It should ultimately be possible to establish a national educational achievement indicator to rival Scholastic Aptitude Test scores--an indicator with the simplicity of the Dow Jones average or the Consumer Price Index. Nine figures illustrate the discussion. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED336420

A Call for a New National Norming Methodology

Glynn Ligon, Ph.D., Evangelina Mangino, Ph.D.

Austin Independent School District
Office of Research and Evaluation
Austin, Texas

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

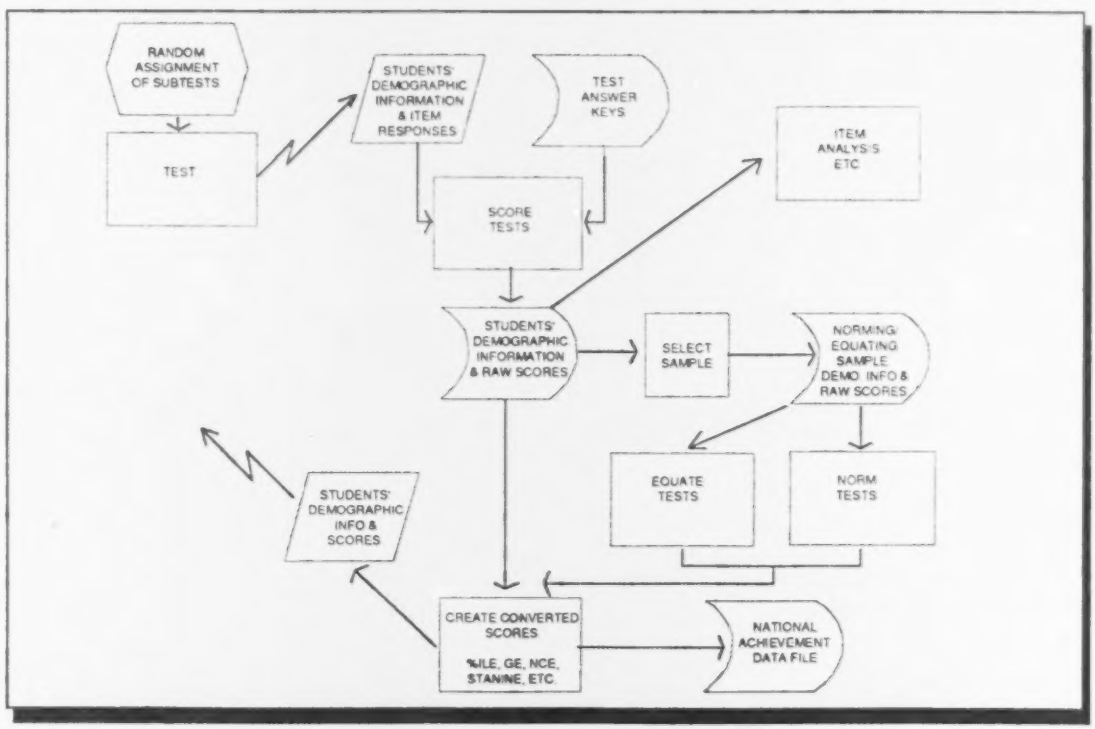
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

GLYNN LIGON

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."



Paper presented at the Annual Meeting of the
American Educational Research Association
Chicago, April, 1991

ORE Publication Number 90.27

TM017217

BEST COPY AVAILABLE

If the U.S. Government Can Tell Us the Consumer Price Index for February in March, Why Do We Have to Wait Years for True National Norms?

OVERVIEW

Testing practitioners, researchers, law makers, parents, counselors, and test publishers are among those who are dissatisfied with national norms for standardized achievement tests. Educators are tiring of being accused of purposely using old test norms to deceive the public into believing that the public schools are doing a better job than they really are. "How Public Educators Cheat on Standardized Tests" (Cannell, 1989) is an fortuitous blundering by a crusader onto some very real shortcomings of standardized test norms and how educators misrepresent them in their reporting. "Does 'nationally' normed really mean nationally?" (Baglin, 1980) is a carefully constructed dissection of a basic flaw in national normings—the over influence of users in norming samples. Test publishers tried to deal with Baglin's criticism by offering incentives other than credit towards the purchase of their own tests or texts to schools for participation in norming. Test publishers tried to deal with Cannell's issues by creating annual user norms. Neither solved the problem.

This paper reviews the issues related to achieving adequate national norms and proposes a new methodology that would work to provide a true measure of national achievement levels on an annual basis and within a timeframe that makes reporting results in current-year norms possible. Yes, this sounds too good to be true, but in reality, advances in computer technology and a few simple lessons learned from, of all sources, the Federal government make this new methodology achievable in the very short term.

A review of recent articles and publications identified why current norming methodology is flawed, what is required for an ideal annual norming, and what is now practical for a functional annual norming.

Why are our current test norms inadequate?

1. They are old the first time they are used.
2. A true national normative sample is virtually impossible to achieve.
3. Annual norms are overly influenced by users with curricula matched to the test objectives.
4. Students taking the tests in the norming sample are unmotivated.
5. Test administrators in the norming sample may or may not follow the standardized procedures closely.
6. Students eligible for the norming may be included or excluded using different rules than in the actual testing; makeups may or may not be administered similarly.
7. Tests are normed only once or twice during a school year requiring interpolation for missing months.
8. Test levels are normed within a limited range of grade levels requiring extrapolation for vertical scaling.

What would be ideal?

A national normative sample with full participation by motivated students, with the same rules for inclusion or exclusion and make-up testing; students tested under true standardized conditions—each month of the school year, every year—in a wide range of grade levels for each test level.

Impossible. However, there is a compromise alternative that could work. Statistical methodology and technology combine now to offer an alternative for a new national norming process that would create an annual national norm published soon after test users complete their regular, annual testing; supplemented by a small national normative sample and adjusted by a factor to reduce the influence of the users. This approach must also include almost immediate turnaround of the current-year norms to allow schools to report their annual scores using those norms.

This paper details how this new approach can work. How schools can participate in this annual, or more frequent

norming, over an electronic network that will allow districts and states to contribute to the norming and receive new norms tables for reporting local scores within a reasonable time period after testing. After all, the national statistics for unemployment, inflation, gross national product, etc. by no means take as long as test publishers now take to release national norms. This paper contrasts the procedures used to gather and report these national statistics with the proposal for a new national norming system.

Schools which are already giving a test would transmit their data electronically to a central location; schools selected in the national normative sample would test and transmit their data. Using appropriate statistical combinations and adjustments, a current-year norms table would be produced and transmitted to local schools; schools could then convert their scores to the current year percentiles and other derived scores for reporting.

Many sampling issues currently make participating in a national norming sample difficult for schools. Publishers typically want students to take an entire battery of tests in order to have a realistic testing situation and in order to calculate composite and total scores. However, a compromise of this procedure could provide current norms for individual tests if we are willing to sacrifice having norms for every total and composite score every year. Individual students could take a single test during the administration of their regular achievement testing, rather than the entire battery. Indeed, this adds to the potential for testing fatigue, and this must be considered in the scheduling by individual districts.

Unfortunately, this type of norming procedure would undoubtedly stretch the planning and computer programming resources of test publishers. Therefore, we must entertain the possibility of creating a national center for test norming—possibly a joint venture among test publishers or a governmental center supported by public funds. Yes, public funds. After all, the need for current and comparable norms is a matter of public interest that apparently cannot be met with the resources available to individual districts or even to individual publishers.

What is the theoretical impact of biased national norms?

Five types of norms have been identified—annual true national norms, point-in-time true national norms, point-in-time user-influenced norms, annual user norms, and annual user-influenced norms. The differences between the scores that would be reported given the use of each of five types of norms will be described. The resulting averages that would be reported for a typical school district under various combinations of a local district's improving, staying the same, or declining; and the national average improving, staying the same, or declining over a six-year period are discussed.

A Call for a New National Norming Methodology

A skeptic might look at this proposal and think that the mechanics of annual norming are too cumbersome to be accomplished quickly; however, current computer technology and a little creative, advance programming can indeed create a system for generating national norms within a reasonable time of the testing.

This is a fascinating time to work in the achievement testing industry in the U.S. While in many states such as Texas, tests are very popular with the legislators, the term authentic assessment is buzzing around everywhere. The advocates of authentic assessments as replacements for traditional multiple choice tests have not yet produced a saleable product that can be adopted by a state, so most of the action in the authentic assessment arena is at the local or national level.

When, in 1980, Roger Baglin challenged the sampling used for national norming, publishers agreed that their test users and their text users made up the preponderance of their national norming samples. However, it took an aspiring psychiatrist, John Cannell, to challenge how every state in the nation could be above those national averages. Some of us thought that the testing establishment would have to make major changes and admit that we were sliding along taking advantage of a few psychometric loopholes to

look good; however, if you saw the fall 1990 issue of Educational Measurement Issues and Practice, you saw a great article by Lorrie Shepard who cited Phillips and Finn, Drahozal and Frisbie, Lenke and Keene, Williams, Qualls-Payne, and Stonehill who all to some degree combined to prove that indeed it is possible for every state to be above the national average.

We have been fascinated by the lack of suggestions of how to solve the real problem—the lack of true annual norms for our standardized tests. In fact, we are fascinated by the apparent fact that throwing out multiple choice tests and embracing alternative assessments is much more popular at this time than is the notion of improving the multiple choice tests.

The bottom line though, is that the testing establishment has done precious little since Cannell began challenging us to respond—other than to attempt to produce annual user norms. Our conclusion after studying annual user norms is that they are better than most people think they are, and we will discuss that in more detail later.

The major impact we see coming from Cannell's criticism has been the movement toward authentic assessments. H.D. Hoover, author of the Iowa Tests of Basic Skills, has a way of simplifying complex issues. At the annual meeting of the Southwest Educational Research Association in January Hoover quoted his favorite definition of authentic assessment as anything other than a standardized test.

President Bush's Education Policy Advisory Committee is planning a new type of national examination that could be in place in the next few years (Rothman, 1990 and 1991). The National Center on Education and the Economy has called for a national standard for all students based upon a series of assessments. They have joined with the Learning Research and Development Center at the University of Pittsburgh to secure $2.5 million in grants to have in place an exam system by the year 2000.

Our reaction to this flurry of activity to abandon the traditional standardized achievement test is—full speed ahead, see what you can do. However, we have reservations about their chances for success. The main reservation is that we strongly suspect that some of the criticism of standardized tests comes from persons who want to pull out from under the burden of accountability. Changing to a different—even better—style of accountability is not what those critics want—they want an end to testing of any kind. Other critics have a sincere desire to make assessment duplicate real applications as much as possible. One of the most troubling predictions about authentic assessments is that, at least initially, the performance gap between ethnic groups will be wider than on standardized tests.

Secondly, every alternative we have heard so far for authentic assessment takes more time, costs more money, and is less reliable across locations than what we have now with multiple choice, standardized tests.

Finally, these national efforts must first solve the issue of local control over curriculum. Just like NAEP, a national effort must negotiate a consensus as to what should be measured. This may not be impossible, but it will be difficult and require some time.

In the meantime, the authors are unwilling to abandon a system of standardized achievement testing that has served us reasonably well, but is suffering from a midlife crisis in norming. We would prefer to solve the norming crisis.

What's wrong with current norms and accountability in education?

"The condition of education" is a phrase we often hear from people from all walks of life, from casual conversations to technical journals and government official reports. Contradictory reports are published almost daily. Do we really know what is happening to the achievement level of the children in this country? Are the schools telling the truth? Do college entrance exams and achievement test scores really indicate trends in achievement? These two indicators seem to tell two contradicting stories. SAT scores have not improved overall—although the percentage of high school graduates taking the test continues to go up, allowing students other than the cream of the crop the opportunity to go to college. Achievement scores, as Cannell indirectly points out, are going up. If using old norms for interpreting current achievement makes schools look better than if they used new norms, it follows that norms are tougher now because overall scores have increased.

The fact that public education, which requires a yearly national investment of over two billion dollars, has flawed national accountability systems in these 'high-tech' times is inconceivable. In the context of national expenditure by all governments, education represents approximately 14%, compared with health and social security (19%), social welfare (16%), transportation (5%), defense (15%), and all other areas of government (31%).

In light of this, our interest in education should be twofold, as an economic investment and as the future of our youth. We need to know how well specific programs work, but we also need much better general accountability for our tax dollar investment.

Other areas of government, particularly those more directly involved with the economic issues, have developed sophisticated means of gathering and analyzing data to obtain trends and indicators that allow decision makers to steer the nation toward progress and a more desirable future.

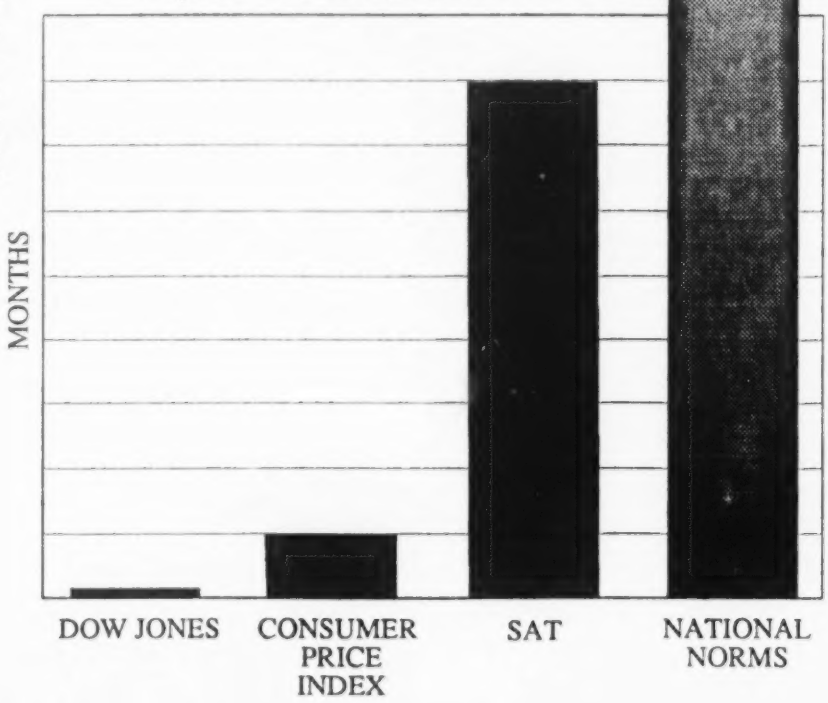
The Consumer Price Index is calculated monthly. The data are collected through the ongoing point-of-purchase Consumer Expenditure Survey. This survey is conducted through visits or calls by trained field representatives from the Bureau of Labor Statistics. The survey includes approximately 21,000 retail and service establishments, 40,000 tenants, and 20,000 owner-occupied housing units in 91 urban areas. Consumer Price Index data are issued every month about three weeks following collection. The Consumer Price Index Detailed Report, available about three weeks after the initial release, provides detailed indexes and a monthly analysis of U.S. price movements. There is also a semiannual report of recent price movements as well as long-term trends.

The Dow Jones Average, on the other hand, is updated daily based on ongoing electronic transmission of data. The price of each particular stock is updated continuously throughout the day, and this information is available worldwide. An investor has access to this information throughout each work day and can make decisions on wether to buy or sell stock.

In education, however, not only is the picture quite nebulous, but it takes a very long time to come out. Indicators of how the nation is doing in the educational arena take at best six months at the college entrance exam level. SAT scores published in September include scores of students tested in March. Approximately one and a half year to two years pass at the elementary and secondary school level before we see test results based on "current" national norms. National norms provided by the test publishing companies take one year to be calculated and a year and a half to two years before the norms are used by local and state education agencies to interpret test data.

In addition to the delayed availability of national norms, the norms are not true national norms. Publishers struggle to get random stratified samples that are representative of students in the country but often they must settle for their third, fourth, or fifth choices of schools because there is not much cooperation from school districts to participate in norming studies. The norms derived this way often are

Figure 1
Periodicity of Reports
Lapse Between Data Collection and Report



based in groups that are users of the test series and/or users of the textbooks and support materials published by the publishing company that publishes the test being normed.

The trend among test publishing companies is to produce user norms annually or biannually, but, as mentioned earlier, the norms do not become available until much after the test is administered, and by the time school districts can actually use them, they are already old norms. Also, publishers only report the norms for the test or tests they publish. Currently, studies equating the different achievement tests are not available. Each test is like a piece of a puzzle and the pieces do not fit together.

Over $80 million have been suggested for the next phase of NAEP and there are plans to develop a national test that would include achievement test scores as well as portfolios and other performance indicators. Why not take advantage of an effort that is already in place and working relatively well? It can work a lot better for the districts using it, as well as to provide a national educational indicator that is timely for making decisions that would affect education at the local and national levels.

What's good about current national norms and standardized tests?

- ❖ They provide a national context for judging local performance.
- ❖ The norms are based on very large samples that meet many of the assumptions required.
- ❖ They are relatively quick and inexpensive to administer.
- ❖ They apparently work to some degree because we continue to use them, Chapter 1 requires them, and the alternatives have yet to be proven.

Throughout the national debate on the influence of old norms and users on norms, it has been a struggle to understand exactly how all these factors work together to give districts either an advantage or a disadvantage. To help illustrate those dynamics, several theoretical situations have been developed.

Taking a very conservative approach, the following assumptions were defined.

- ✦ If one assumes that a test user nets an advantage each year through familiarity with the test, then the minimum advantage gained is +1 percentile annually.
- ✦ If one assumes that the user of a publisher's text nets an advantage each year through familiarity with the test contents, then the minimum advantage gained is +1 percentile annually.
- ✦ If a new test were to be used each year, then these advantages for a user would be nullified.
- ✦ A user in a user-influenced norm group who is at the true 50th percentile nationally would outscore nonusers in the norm group and tie other users, thus netting some average of the two 1 percentile point advantages for users and would score at the 51st percentile rather than the 50th percentile.

Figures 2-4 show the theoretical outcomes from six years of testing for a district using various assumptions about the achievement trends across those years.

Now in our first example, Figure 2, we have a district that is a test user, a text user, is at the national average in year one, maintains that same level of achievement over the six years, while users and nonusers do the same nationally.

The influence of all these factors results in a maximum difference of 11 percentile points in year 6 across the 10 combinations of norm types and tests shown. That is a major influence, using very conservative assumptions. If you believe that an 11 percentile point bias is higher than reality, then maybe those conservative assumptions were not so conservative at that. In that case, the influence of the user factors described in the assumptions would be more minimal than some may think.

Interestingly, the combinations that most closely match the results from an annual true national norm using a new test each year is annual user norms using the same test each year or a new test each year. Both yield a percentile of 50 in year 6.

Figure 2

Achievement Trends: Local Maintaining, National Maintaining

Assumptions

Local School District

- Text User *(+1 %ile each year same test is used)*
- Test User *(+1 %ile each year same test is used)*
- At 50th %ile in Norming Year, True National Norm
- Maintaining Same Achievement Level Annually

Users Nationally

- Average at 50th %ile, True National Norm
- Follow Same Achievement Trend as National Trend

National Achievement

- Maintaining Same Achievement Level Annually

		Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
Annual True National Norm	New Test Each Year	50	50	50	50	50	50
	Same Test Each Year	50	52	54	56	58	60
Point-in-Time True National Norm	New Test Each Year	50	50	50	50	50	50
	Same Test Each Year	50	52	54	56	58	60
Point-in-Time User-Influenced Norm	New Test Each Year	51	51	51	51	51	51
	Same Test Each Year	51	53	55	57	59	61
Annual User Norm	New Test Each Year	50	50	50	50	50	50
	Same Test Each Year	51	50	50	50	50	50
Annual User-Influenced Norms	New Test Each Year	51	51	51	51	51	51
	Same Test Each Year	51	52	53	54	55	56

Our second example in Figure 3 differs from the first only in that the local and national averages are going up 2 percentile points annually. In this context, an annual user norm is also as accurate as a true national norm when a new test is administered each year. Few districts do that, so the more important finding is that the annual user norm tracks an annual true national norm whether or not a new test is given each year. The most distorted view comes from a point-in-time user-influenced norm—which is exactly what most districts use for reporting.

Our third example in Figure 4 illustrates the situation when local and national achievement is declining at a 2 percentile point per year rate. The dynamics are that the local user

Figure 3

Achievement Trends: Local Up, National Up

Assumptions

Local School District

- Text User *(+1 %ile each year same test is used)*
- Test User *(+1 %ile each year same test is used)*
- At 50th %ile in Norming Year, True National Norm
- Gaining 2 %ile points from Norm Year Annually

Users Nationally

- Average at 50th %ile, True National Norm
- Follow Same Achievement Trend as National Trend

National Achievement

- Gaining 2 %ile points from Norm Year Annually

		Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
Annual True National Norm	New Test Each Year	50	50	50	50	50	50
	Same Test Each Year	50	52	54	56	58	60
Point-in-Time True National Norm	New Test Each Year	50	52	54	56	58	60
	Same Test Each Year	50	54	58	62	66	70
Point-in-Time User-Influenced Norm	New Test Each Year	51	53	55	57	59	61
	Same Test Each Year	51	55	59	63	67	71
Annual User Norm	New Test Each Year	50	50	50	50	50	50
	Same Test Each Year	51	50	50	50	50	50
Annual User-Influenced Norms	New Test Each Year	51	51	51	51	51	51
	Same Test Each Year	51	52	53	54	55	56

district is advantaged by the use of the same test each year, the losses over time are masked along with the fact that local achievement is tracking the national average.

If we were to write a handbook of how to use national norms to rank high on standardized tests, what would we say?

If national achievement is remaining stable, and you are a text/test user, then there is no real advantage or disadvantage in annual user norms. The big advantage comes when users give the same test annually and compare themselves to the old, point-in-time national norm—either a true national norm or a user-influenced one.

If national achievement is going up, a point-in-time, user-influenced norm is the most advantageous. This confirms that the test publishers have been on to a great marketing strategy over the last two decades.

Figure 4

Achievement Trends: Local Down, National Down

Assumptions

Local School District

- Text User *(+1 %ile each year same test is used)*
- Test User *(+1 %ile each year same test is used)*
- At 50th %ile in Norming Year, True National Norm
- Losing 2 %ile points from Norm Year Annually

Users Nationally

- Average at 50th %ile, True National Norm
- Follow Same Achievement Trend as National Trend

National Achievement

- Losing 2 %ile points from Norm Year Annually

		Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
Annual True National Norm	New Test Each Year	50	50	50	50	50	50
	Same Test Each Year	50	52	54	56	58	60
Point-in-Time True National Norm	New Test Each Year	50	48	46	44	42	40
	Same Test Each Year	50	50	50	50	50	50
Point-in-Time User-Influenced Norm	New Test Each Year	51	49	47	45	43	41
	Same Test Each Year	51	51	51	51	51	51
Annual User Norm	New Test Each Year	50	50	50	50	50	50
	Same Test Each Year	51	50	50	50	50	50
Annual User-Influenced Norms	New Test Each Year	51	51	51	51	51	51
	Same Test Each Year	51	52	53	54	55	56

If national achievement is going down, an annual true national norm or a user norm would keep a typical district at the national average. The worst case would be to use the old point-in-time national norm, because your district would fall behind the artificially stable national norm. However, if the downward trend is equivalent to the advantage gained from being a test/text user who gives the same test annually, the loss in achievement is effectively masked.

The point of all this is to confirm that the best norm is an annual norm—based either on a true national sample or on a user sample (if you are a user also).

However, we need to acknowledge a pitfall of an annual norm. Imagine that your district is soaring ahead and making tremendous gains. At the same time, the national average is going up. You have told your superintendent and board of trustees about these great gains, and they have asked you to provide a graph to illustrate them.

Your district's longitudinal graph would look like Figure 5—flat. This is the similar to the dilemma we face when parents and teachers of gifted students complain that their students score at the 99th percentile every year and show no progress—they never can go up.

What this confirms is that there is a legitimacy in using a point-in-time national norm. As Figure 6 shows, setting a baseline year for future comparisons provides an indication of the trend over time. The problem we have had with critics like Cannell is that we have not communicated that legitimacy and have given the impression that we are making comparisons to current national achievement levels. If we were to conduct a local survey of the price of a gallon of gas, then compare that to the national average price from two years ago and claim that local prices are cheaper, most citizens off the street would realize we were not operating with a full tank of gas. However, that is exactly what we do when we use point-in-time test norms.

Comparing to a base year is legitimate; however, we must communicate to the audiences that is what we are doing.

Figure 5

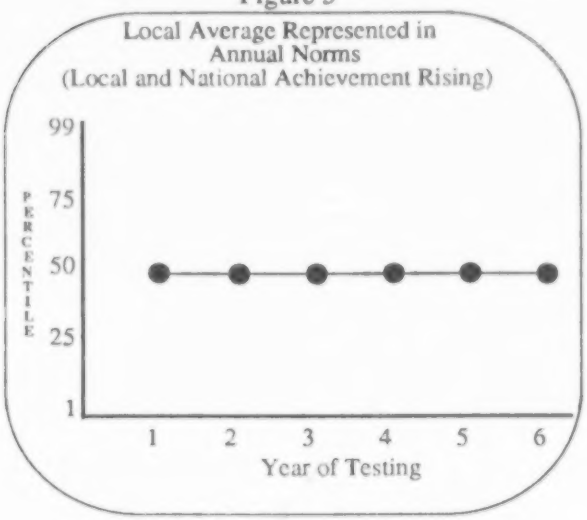
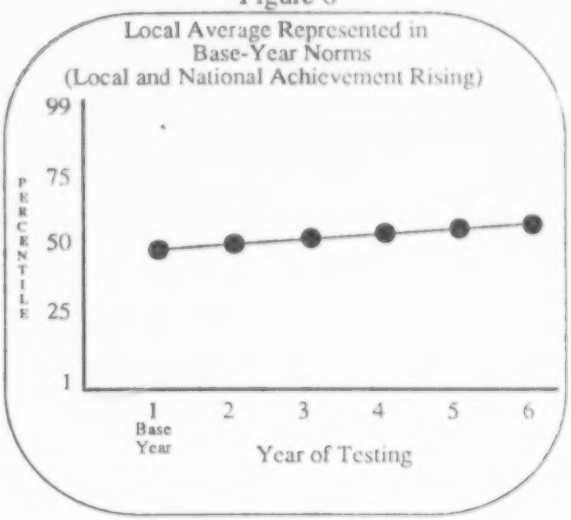


Figure 6



What is required for a true and timely national norm?

- ❖ Participation by a sample that truly represents students across the nation
- ❖ Speedy results

This does not mean results the next year, but results fast enough to use in the current year's report. Yes, if we test in April, we should be able to report those scores in current-year percentiles during the summer.

Why do we think this can be accomplished?

- ✦ Scoring of tests by districts and scoring services has improved dramatically. Most districts get 2 - 4 week turnaround from scoring services, and districts with in-house scoring accomplish the task even quicker. In Austin, the testing staff collects answer sheets by the Friday after testing and delivers results to the schools on Monday morning.
- ✦ To get the scores to a central location for calculating the norms, we now have national electronic networks.

An example of a somewhat similar communications issue that is being addressed through electronic networking is the National Center for Educational Statistics (NCES) Interstate Student Record Transfer System (ISRTS), which is currently in a pilot stage. This system is setting a national standard for formatting student information and sending it across networks between public schools, between public schools and institutions of higher education, and between institutions of higher education. The goals include decreasing the time required to transmit records and reducing the costs for sending paper records that must be reentered into the new school's computer files.

- With today's computers and with some advanced programming, the number crunching required can be accomplished quicker.

- Those same electronic networks can distribute the new norms tables back to school districts.

In Roger Baglin's 1980 paper on self-selection bias in national norms, he raised the notion of a joint norming effort by the test publishers. There are restraint-of-trade laws that would kill such cooperation. However, the cost for a national clearinghouse funded by the federal government would be much less than the cost for NAEP or any of the other national efforts being planned.

There would be less costs for test development, less for additional tests to be printed, and less for scoring services. With IRT and other equating methods, local choice of tests could continue, and we could still make comparisons across states. The problems encountered in the 1970 anchor study could not be ignored, but could be addressed.

National Achievement Indicator and Norming Program

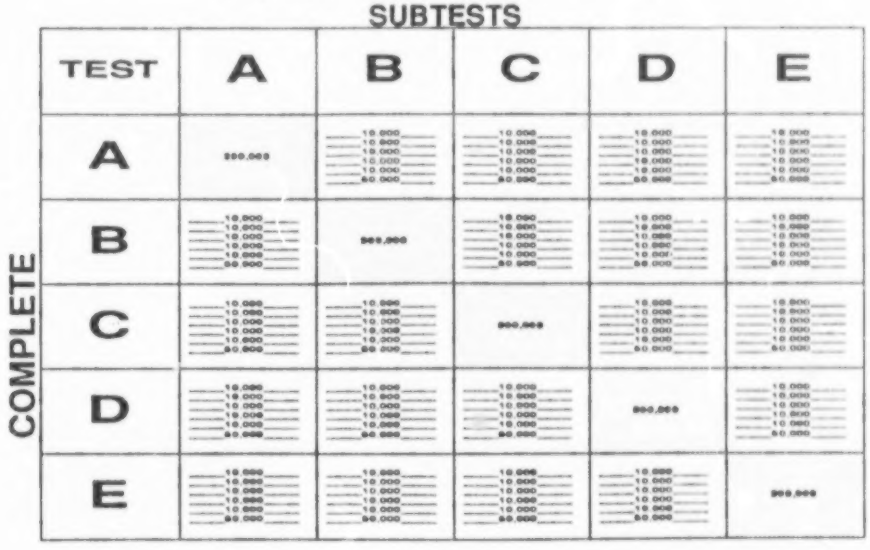
The proposed solution to the timely norming problem and the creation of a national educational indicator could be obtained by centralizing the creation of annual norms for the major standardized achievement tests. This could be done through the use of electronic transmission of data and the voluntary cooperation of the districts using these tests.

Cooperation from school districts would be secured by the fact that they would get current national norms within weeks of transmitting their raw data. Currently, districts are reluctant to participate in norming studies because that implies additional testing--either a new version of the test they are currently using or a complete battery of a test they are not using and do not intend to adopt in the future. With this system, the districts would be required to administer one extra subtest to each student in addition to the test they normally administer.

Currently, a national norming study requires 100,000 to 200,000 students taking the test (including grades K-12). With the proposed program, a norming group could be much larger, as illustrated in Figure 7. An additional subtest would be assigned by the norming center based on stratified sampling methods that would distribute the subtests in such a way that the sample size and makeup would provide a valid norming group.

In the example presented in Figure 7, the norming center would select 200,000 test users to be included in the norming sample. Four sub-sets of 50,000 students each would be administering subtests from four other tests. Within each subset of students, five groups of 10,000 students each would be administering different subtests of the assigned test.

Figure 7
Sampling Paradigm for National Norms

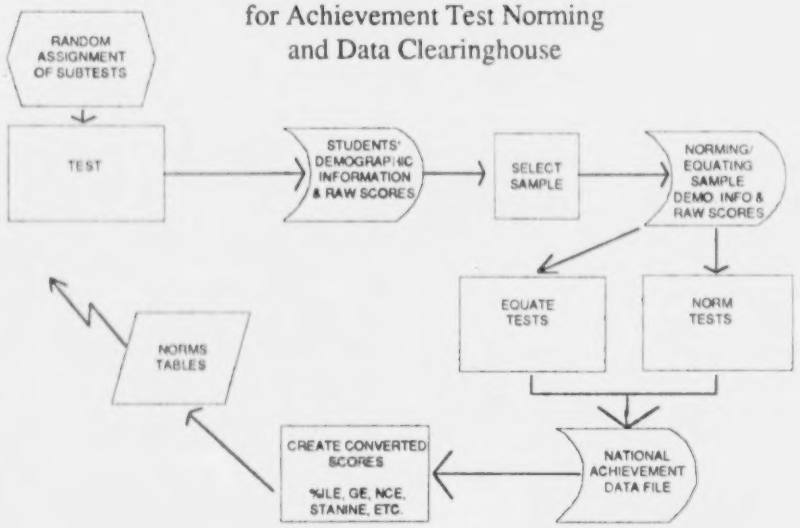


BEST COPY AVAILABLE

The additional subtest would be used for equating purposes, thus linking together all the tests to create norms that are not user influenced. One of the sacrifices that could be made to make the system functional is multiyear norms for total and composite scores. Most tests combine various subtests into these composites in order to give users fewer scores to deal with in reporting and selecting students for programs. In reality those composites require that students in the norming sample take all tests included, because the composites are more than the mere arithmetic average of subtests. However, with past experience on our side, a national center could estimate composite score norms from individual test norms. This would allow for a testing schedule that requires individual students to take fewer subtests each year. Periodic testing with a full battery could be done at less cost and less disruption of instruction.

After test administration in spring and fall, districts would process their tests in the way they are accustomed to (either in-house scanning, using a scanning service, or the services offered by the test publishers). The data would be electronically transmitted to the norming center and, within approximately a month, they would receive the current national norms with which to interpret their current year's achievement.

Figure 8
Flow of Information: National Center
for Achievement Test Norming
and Data Clearinghouse



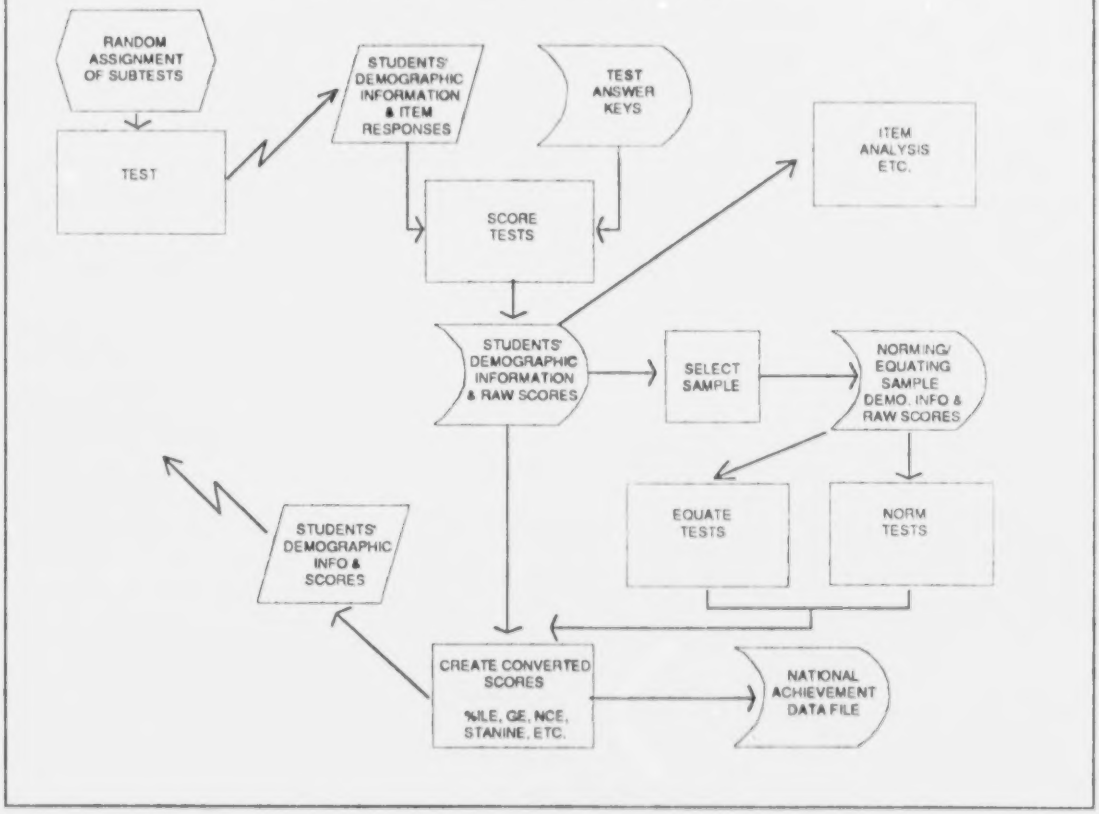
The notion of a national achievement indicator to rival SAT/ACT scores is a dream of NAEP and the newer national testing movements. However, the simplicity of the Dow Jones' Average, the Consumer Price Index, or the SAT/ACT averages is a real challenge. A national achievement indicator could be a composite across all tested areas or merely be two representative areas such as reading and mathematics. Remember, simplicity is the key, so having separate numbers for all subject areas detracts from the utility of the indicator. A simple option is to create a scale with 100 being the national average for a base year and achievement test scores being converted to this scale annually.

There are two options for the norming center. One is the most conservative (Figure 8), having the center be only a norming center that collects raw scores and demographic data and returns tables to the districts for score interpretation.

In the second option (Figure 9), the center is a national scoring service and norming center, where the districts would transmit item responses and demographic information for each student. The center would score the tests, transmit the item responses to the test publishers for research purposes, and electronically transmit students' converted scores back to the districts within a few weeks.

The first option has the advantage of a lower cost and simplicity of operation.

Figure 9
Flow of Information:
National Center for Achievement Test Scoring,
Norming Center, and Data Clearinghouse



Conclusion

School systems are spending millions of dollars for norm- referenced achievement tests, but we do not have true national norms. Now is the time to take advantage of existing technology to give us timely, true national norms.

Together, we should work toward the day when a state legislator, or a governor, or even the education president wakes up in the morning and checks the paper for that day's Dow Jones' Average, the Consumer Price Index, and of course, the national achievement level and says, "Wow, it was a good week for public education."

BEST COPY AVAILABLE

References

Baglin, R.F. (1981). Does "nationally" normed really mean nationally? Journal of Educational Measurement, 18, 97-107.

Cannell, J.J. (1987). Nationally normed elementary achievement testing in America's public schools: How all fifty states are above the national average. Daniels, WV: Friends for Education.

Cannell, J.J. (1989). How public educators cheat on standardized achievement tests. Albuquerque, NM: Friends for Education.

Drahozal, E.C. and Frisbie, D.A. (1988). Riverside comments on Friends for Education Report. Educational Measurement: Issues and Practice, 7, 12-16.

Hertzberger, M.P. and Beckman, B.A. (1989). Business cycle indicators: Revised composite indexes. Survey of Current Business, 23-28.

Rothman, R. (1990). Two groups laying plans to develop national exams. Education Week, X, 4, 1 and 14.

Rothman, R. (1991). Promise, pitfalls seen in creating national exams. Education Week, X, 19, 1 and 18.

Shepard, L.A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? Educational Measurement: Issues and Practice, 9, 15-21.